Minimal clade size in the Bolthausen-Sznitman coalescent

F. Freund, A. Siri-Jégousse

March 7, 2013

Abstract

This article shows the asymptotics of distribution and moments of the size X_n of the minimal clade of a randomly chosen individual in a Bolthausen-Sznitman n-coalescent for $n \to \infty$. The Bolthausen-Sznitman n-coalescent is a Markov process taking states in the set of partitions of $\{1, \dots, n\}$, where $1, \dots, n$ are referred to as individuals. The minimal clade of an individual is the equivalence class the individual is in at the time of the first coalescence event this individual participates in.
The main tool used is the connection of the Bolthausen-Sznitman n-coalescent with random recursive trees introduced by Goldschmidt and Martin (see [16]). This connection shows that $X_n - 1$ is distributed as the number M_n of all individuals not in the equivalence class of individual 1 shortly before the time of the last coalescence event. Both functionals are distributed like the size RT_{n-1} of an uniformly chosen table in a standard Chinese restaurant process with $n-1$ customers.We give exact formulae for these distributions.
Using the asymptotics of M_n shown by Goldschmidt and Martin in [16], we see $(\log n)^{-1} \log X_n$ converges in distribution to the uniform distribution on [0,1] for $n \to \infty$.
We provide the complimentary information that $\frac{\log n}{n^k} E(X_n^k) \to \frac{1}{k}$ for $n \to \infty$, which is also true for M_n and RT_n.

Keywords: minimal clade size, Bolthausen-Sznitman n-coalescent, Chinese restaurant process

AMS 2010 Mathematics Subject Classification: Primary 60C05; Secondary 05C80, 60G09, 60F05, 60J27, 92D25;

1 Introduction

The Bolthausen-Sznitman n-coalescent is a time-homogeneous Markov process $(\Pi_t^{(n)})_{t \geq 0}$ whose state space is the set of partitions of $\{1, \dots, n\}$. The only possible transitions in this process are those in which several blocks of a partition are merged (or coalesced) into one new block. Only one new block can be formed in a transition (no simultaneous mergers). Each k-tuple of b present blocks is merged to a new block at rate $\frac{(k-2)!(b-k)!}{(b-1)!}$. The Bolthausen-Sznitman n-coalescent is a member of the Λ-n-coalescent family (which were introduced independently by Sagitov [24] and Pitman [22]). A Λ-n-coalescent is again a time-homogeneous, continuous-time Markov process whose state space is the set of partitions of $\{1, \dots, n\}$. The possible transitions are again mergers of multiple blocks into a

new one. Each merger of k blocks among b happens with rate

$$\int_{[0,1]} x^{k-2}(1-x)^{n-k}\Lambda(dx)$$

for a finite measure Λ on $[0,1]$. Note that the Bolthausen-Sznitman coalescent has $\Lambda = U_{[0,1]}$, the uniform distribution on $[0,1]$. Each Λ-n-coalescent can be represented as a random tree with n leaves $\{1,\dots,n\}$ and random branch lengths by representing each merger as an internal node in the tree (the branch lengths are then the waiting times for the mergers, time is measured starting from the leaves). Also note that a Λ-n-coalescent at time t forms a random exchangeable partition of $\{1,\dots,n\}$.

The Bolthausen-Sznitman n-coalescent was introduced by Bolthausen and Sznitman in 1998 (see [4]). It has connections to population genetics and physics. In mathematical physics, it appears in the context of spin glasses (see [4] and [5]). It also seems to be a suitable model for the genealogy of a sample of n alleles/genes/haplotypes in several models for selection in population genetics (see [8], [9], [2], [19], [12] see also the survey [7]). Note that this is in contrast to the standard model for a genealogical tree of such a sample which is Kingman's n-coalescent ($\Lambda = \delta_0$, only 2 merger at a time, introduced in [21]). Also note that due to the interpretation of the Bolthausen-Sznitman n-coalescent as a genealogical tree, we refer to $\{1,\dots,n\}$ as individuals.

Here, we focus on the Bolthausen-Sznitman n-coalescent as a model for a genealogical tree which depicts the ancestry of n alleles sampled at a genetic locus. Since the genealogical tree often is endowed with a mutation structure which is interpreted under the infinitely-many sites model, we assume a locus consisting of many nucleotide sites, for example a gene. Different alleles can thus also be seen as different haplotypes at the according sites. One important information coded in the genealogy is the relatedness of an allele randomly chosen from the sample to the rest of the sample. There are two functionals/statistics of the genealogical tree which transport complementary information about this relatedness. The first functional is the length E_n of an external branch chosen at random from the n external branches associated with the leaves $\{1,\dots,n\}$ of the tree, introduced by Fu and Li in [15]. E_n gives the time that the chosen allele has to evolve independently of the rest of the sample (e.g., by mutation). This gives a measure of the genetic uniqueness of this allele relative to the rest of the sample. The second functional is the size X_n of the minimal clade containing the randomly chosen allele, introduced by Blum and François in [6]. The minimal clade can be defined in different, yet equivalent ways: The minimal clade is

- the equivalence class that contains the (randomly chosen) allele $i \in \{1,\dots,n\}$ at the first time i was merged,
- all leaves of the subtree rooted at the most recent ancestor of allele i,
- all descendants of the most recent ancestor of allele i.

The minimal clade can also be seen as the smallest family containing i. The size of the minimal clade gives the complementary information how many individuals

share the genealogy with allele i "after" time E_n (note that since we measure time from leaves to root, "after" E_n actually means further back in time).
The external branch length is already analyzed well for several Λ-n-coalescents in the literature. Its distribution follows a recursion and its asymptotics for sample size $n \to \infty$ are known for various Λ-n-coalescents (see [14], [10], [6], [17], [13]). For the minimal clade size, though, only results for Kingman's n-coalescent ($\Lambda = \delta_0$, only 2 merger at a time) are known (including asymptotics for $n \to \infty$, see [6]).
The purpose of this paper is to analyze the distribution of the minimal clade size X_n in the case of the Bolthausen-Sznitman n-coalescent and its asymptotics for sample size $n \to \infty$. We will exploit the construction of the Bolthausen-Sznitman n-coalescent using a random recursive tree introduced by Goldschmidt and Martin (see [16]) to prove our results. First, we observe that this construction yields that the process describing the set of relatives of a randomly chosen individual in the Bolthausen-Sznitman n-coalescent process (which is its equivalence class without the individual itself) is equal in law to the time-reversed process describing the set of non-relatives of the chosen individual (all individuals in different equivalence classes than the chosen individual). This shows that the minimal clade size actually is distributed as the sum M_n of the sizes of all blocks not containing 1 which participate in the last collision in the n-coalescent. Convergence in distribution of properly scaled M_n for $n \to \infty$ was shown already by Goldschmidt and Martin in [16] and thus the same asymptotic behavior holds for X_n, namely $(\log n)^{-1} \log X_n$ converges in distribution to the uniform distribution on [0,1].
Note that due to the connection between the random recursive tree and the standard Chinese Restaurant process, we observe that $X_n - 1$ and M_n are distributed as the size of a uniformly chosen table (not chosen by a size-biased pick!) in the Chinese restaurant process (again for M_n in accordance to [16]). This allows us to give several formulae for the exact distribution of X_n. Using these, we show that $\frac{\log n}{n^k} E(X_n^k) \to \frac{1}{k}$ for $n \to \infty$, which gives complementary information to the weak convergence result.

2 Minimal clade size in the Bolthausen-Sznitman n-coalescent

Set $[n] := \{1, \dots, n\}$ and $[n]_0 := \{0, \dots, n\}$. For a partition η of $[n]$, let $C_i(\eta)$ denote the equivalence class of $i \in [n]$ and $|C_i(\eta)|$ its size. Let $(\Pi_t^{(n)})_{t \geq 0}$ be a Λ-n-coalescent. Since we want to look at the minimal clade size of a randomly chosen allele in the sample whose genealogy is given by $(\Pi_t^{(n)})_{t \geq 0}$, define I as a uniform pick from $[n]$ independent of the n-coalescent. Now, first define the length of a randomly chosen external branch (associated with the randomly chosen $I \in [n]$) by

$$E_n := \inf\{t \geq 0, C_I(\Pi_t^{(n)}) \neq \{I\}\}.$$

Now we define the size of the minimal clade of the randomly chosen allele I as

$$X_n := |C_I(\Pi^{(n)}_{E_n})|. \tag{1}$$

Note that, due to exchangeability, we don't change the distributions of E_n and X_n if we assume $I = 1$. Also note that due to the interpretation of a n-coalescent as a genealogical tree, we refer to $\{1, \dots, n\}$ as individuals.
From now on, we will abbreviate the size of the minimal clade of $I = 1$ with minimal clade size.
The minimal clade of individual 1 is the size of the equivalence class of 1 at the first coalescence event that the individual participates in. In [16], Goldschmidt and Martin have analysed the behavior of the total mass M_n of the equivalence classes not containing 1 at the last coalescence event in the Bolthausen-Sznitman n-coalescent (see [16, Thm. 3.1]). Note that M_n can also be written as $n - |C_1(\Pi^{(n)}_{\tau_n -})|$, where τ_n is the waiting time for the last coalescence event. Both X_n and M_n are functionals of the equivalence class of 1 in the Bolthausen-Sznitman n-coalescent at different times. Thus, it's interesting how the equivalence class of 1 changes over time. It will only grow by merging with other equivalence classes at coalescence times, but not necessarily at all coalescence times. We define $S^{(n)}_i$ as the equivalence class of 1 after the ith merging event which 1 participates in. What are the properties of $(S^{(n)}_i)_{i \in [\kappa_n]_0}$, where κ_n is the number of merging events 1 participates in? The results from [16] answer this question. There, the authors show a construction of the Bolthausen-Sznitman n-coalescent by applying a cutting procedure to a random recursive tree and use it, among other questions, to analyse M_n.
We will show in detail that this construction enables us to analyse the behaviour of $S^{(n)}$ and that it can be expressed in terms of a Chinese retaurant process. Note that this is just the line of reasoning from [16]. Let's quickly recall the construction of the Bolthausen-Sznitman n-coalescent from a random recursive tree from [16, Prop. 2.2] as well as the connection to the Chinese restaurant process. Here, we give a simplified version just constructing the jump chain of the n-coalescent.
We start with a random recursive tree with n vertices, i.e. a uniformly distributed random variable on the set of all recursive trees with n vertices $1, \dots, n$ rooted in 1 (here, the branches carry no length information). Now construct the jump chain as follows.

1. Choose an edge at random
2. Cut the tree at this edge. All labels that are in the subtree not containing the root are added to the node of the subtree containing the root which was adjacent to the cut edge.
3. Define a partition by taking the labels at each node of the subtree containing the root. This partition has the same law as the Bolthausen-Sznitman n-coalescent after the first jump.

4. Repeat subsequently steps 1-3 with the subtree containing the root. This leads to partitions which have the same law as the Bolthausen-Sznitman n-coalescent after the 2nd, 3rd, ... jump.

We now come to the connection of the random recursive tree with the Chinese restaurant process.
First, recall that the standard Chinese restaurant process is a sequential construction of a uniform permutation of $[n]$. Imagine a restaurant with tables $1, 2, 3, \ldots$ with infinitely many chairs. n customers $1, \ldots, n$ sit down at the tables after the following rule:

- customer 1 sits at table 1,
- if $i-1$ customers have taken their seat, the ith customer sits with equal probability at one of the following i places:
  - on a chair directly to the left of an already seated customer (possibly between customers),
  - at a previously unoccupied table.

Writing down the customers at each table in seating order, we get the cycles of a uniform random permutation of $[n]$. If we only record the customers at each table, but not the seating order, we get an exchangeable partition of $[n]$ whose distribution is given by Ewens sampling formula. More information on this process can be found in [23, Ch. 3.1]. We will abbreviate a standard Chinese restaurant process with n customers by $CRP(n)$.
A $CRP(n-1)$ can be found in a random recursive tree with n vertices in the following way (see [16, p. 724-725]). We define a subtree of '1' in the random recursive tree as a rooted subtree whose root is adjacent (connected by one edge) to the root '1' of the whole tree. Then the subtrees of '1' form a exchangeable partition of $\{2, \ldots, n\}$ which can be described as a $CRP(n-1)$ with customers labelled $2, \ldots, n$. The following lemma just is a write-up of the line of reasoning from [16, p. 725] and gives a discrete analogon of a part of [22, Cor. 16] (in [16], the line of reasoning presented here is a part of an alternative proof for [22, Cor. 16])

Lemma 2.1. *(practically from Goldschmidt, Martin) Let κ_n be the number of collisions in a Bolthausen-Sznitman n-coalescent individual 1 participates in. For $i \in [\kappa_n]_0$, let $S_i^{(n)}$ be the equivalence class of 1 in the Bolthausen-Sznitman n-coalescent after the ith collision. For a $CRP(n-1)$ with K_{n-1} tables, let $RT_1, \ldots, RT_{K_{n-1}}$ be the tables in random order. Then $S^{(n)} = (S_i^{(n)})_{i \in [\kappa_n]_0}$ is distributed as $(\{1\} \cup \bigcup_{j \in [i]} RT_j)_{i \in [K_{n-1}]}$.*
Moreover, the process $S^{(n)} \setminus \{1\} = (S_i^{(n)} \setminus \{1\})_{i \in [\kappa_n]_0}$ giving the relatives of individual 1 through time is distributed as the time-reversed process $[n] \setminus S^{(n)} = ([n] \setminus S_{\kappa_n - i}^{(n)})_{i \in [\kappa_n]_0}$ giving the nonrelatives of individual 1.

If the Bolthausen-Sznitman n-coalescent is constructed via cutting a random recursive tree, this lemma can be described more graphically: The equivalence

class of 1 grows by adding tables chosen uniformly at random from the Chinese restaurant process with $n-1$ customers given by the subtrees of '1' in the random recursive tree.
Note that we actually chose tables at random not individuals sitting at tables, so we don't make size-biased picks.

Proof. We construct the Bolthausen-Sznitman n-coalescent via cutting a random recursive tree. The equivalence class of 1 is merged with other equivalence classes as soon as an edge adjacent to the root is cut in the random recursive tree. The equivalence class of 1 is then merged with the subtree of '1' which is connected by that edge. Since the edges are chosen at random, this means that a uniformly chosen table of the $CRP(n-1)$ given by the subtrees of '1' is merged with the class of 1. □

Since $X_n - 1 = |S_1^{(n)} \setminus \{1\}|$ and $M_n = |[n] \setminus S^{(n)}_{\kappa_n - 1}|$, Lemma 2.1 shows that $X_n - 1$ and M_n have the same distribution, namely that both are distributed as the size of a uniformly chosen table in a $CRP(n-1)$. This means that the known results for the asymptotics of M_n which are given in [16, Thm. 3.1] are valid for $X_n - 1$ and due to a Slutski argument are also valid for X_n.

Theorem 2.2. *Let $n \in \{2, 3 \ldots\}$. Let X_n be the minimal clade size in the Bolthausen-Sznitman n-coalescent. X_n is distributed on $2, \ldots, n$. X_n is distributed as the size of a randomly chosen table in a $CRP(n-1)$ reduced by 1 and*

$$\frac{\log X_n}{\log n} \to U_{[0,1]}$$

holds in distribution for $n \to \infty$, where $U_{[0,1]}$ is the uniform distribution on $[0,1]$.

Additionally to this result, we give the complementary information of the exact law of X_n and of the first order behaviour of all moments of X_n for $n \to \infty$. For this, we need more knowledge about the distribution of X_n.
Theorem 2.2 states that the distribution of X_n can be expressed in terms of the Chinese restaurant process. We will use this to derive three formulae for the distribution of X_n. Let's recall two possibilities to look at the distribution of customers at tables in a $CRP(n)$. It is well known that this distribution in a $CRP(n)$ is given by the celebrated Ewens sampling formula with mutation parameter $\theta = 1$ (e.g., see [1, eq. 1.3]). We use two different possibilities to look at the Ewens sampling formula in equations (2) and (3). First, we can record how many tables in a $CRP(n)$ have exactly i customers, which we denote by $A_i^{(n)}$, for each $i \in [n]$. Then for $a_1, \ldots, a_n \in [n]_0$ with $\sum_{i \in [n]} i a_i = n$, we have

$$\mathbb{P}(A_1^{(n)} = a_1, \ldots, A_n^{(n)} = a_n) = \prod_{i=1}^{n} \frac{1}{a_i! i^{a_i}}. \tag{2}$$

On the other hand, we can record the probability that certain sets of customers sit at tables $1, 2, \ldots$ (this forms a partition η of $[n]$). The probability that we

find a certain partition η (with blocks ordered by their least element) of $[n]$ with k occupied tables and n_i customers at the ith occupied table is

$$P(CRP(n) = \eta) = \frac{1}{n!} \prod_{i \in [k]} (n_i - 1)!. \tag{3}$$

This leads to several possibilities to express the distribution of X_n.

Lemma 2.3. *Let $n \in \{2, 3, ...\}$. Let X_n be the minimal clade size in a Bolthausen-Sznitman n-coalescent. For $m \in \mathbb{N}$, let $A_i^{(m)}$ be the number of tables with exactly i customers in a $CRP(m)$ and $K_m = \sum_{i \in [n-1]} A_i^{(m)}$ the number of occupied tables. Define $K_0 = 0$ ('empty restaurant') Then for $j \in [n-1]$*

a) Denoting $\Gamma_n = \{a_1, \dots, a_{n-1} \in [n-1]_0, \sum_{i=1}^{n-1} i a_i = n-1\}$

$$\begin{aligned} P(X_n = j+1) &= E\left(\frac{A_j^{(n-1)}}{K_{n-1}}\right) \\ &= \sum_{\Gamma_n} \frac{a_j}{\sum_{i=1}^{n-1} a_i} \prod_{i=1}^{n-1} \frac{1}{a_i! i^{a_i}}. \end{aligned}$$

b) Denoting $\Delta(n,k) = \{n_1, \dots, n_k \in [n], \sum_{i=1}^{k} n_i = n\}$ for $k \leq n$,

$$P(X_n = j+1) = \frac{1}{j} \sum_{k=1}^{n-1-j} \frac{1}{(k+1)!} \sum_{\Delta(n-1-j,k)} \frac{1}{n_1 \cdots n_k}$$

for $j < n-1$ and $P(X_n = n) = \frac{1}{n-1}$.

c) Let $B_1, B_2, \dots$ be independent Bernoulli-distributed random variables with success probability $\frac{1}{i}$ for B_i.

$$\begin{aligned} P(X_n = j+1) &= \frac{1}{j} E\left(\frac{1}{1 + K_{n-1-j}}\right) \\ &= \frac{1}{j} E\left(\frac{1}{1 + \sum_{i=1}^{n-1-j} B_i}\right), \end{aligned}$$

Note that above lemma also holds true for M_n and the size RT_{n-1} of a randomly chosen table in a $CRP(n-1)$ (just replace $j+1$ with j). Also note that this result provides a very rare example where an exact law is obtained for a functional of an exchangeable non-Kingman, non-starshaped n-coalescent.

Proof. Due to Theorem 2.2, we know that $X_n - 1$ is distributed as the size of a randomly chosen table in a $CRP(n-1)$. Given the table counts $A_1^{(n-1)}, \dots, A_{n-1}^{(n-1)}$, the probability that we randomly choose a table with j

customers is $\frac{A_j^{(n-1)}}{\sum_{i=1}^{n-1} A_i^{n-1}} = \frac{A_j^{(n-1)}}{K_{n-1}}$. Summing over the distribution of the table counts given by (2) gives a).
Now look at the partition η of $[n]$ constructed via a $CRP(n-1)$ whose distribution is given by (3). We are interested in the partition not in order of least elements, but in exchangeable order (meaning that if the partition has k blocks, we order them randomly). Let $N_1^{(n-1)}, \dots, N_k^{(n-1)}$ be the table sizes in exchangeable order. By combinatorial arguments (see [23, (2.7)]), we get

$$P(N_1^{(n-1)} = n_1, \dots, N_k^{(n-1)} = n_k) = \binom{n-1}{n_1, \dots, n_k} \frac{1}{k!} \underbrace{\frac{1}{(n-1)!} \prod_{i=1}^{k} (n_i - 1)!}_{\text{prob. of } \eta, \text{ least elements}}$$
$$= \frac{1}{k! \prod_{i=1}^{k} n_i}. \tag{4}$$

The size of a randomly picked table in the CRP is distributed as $N_1^{(n-1)}$. This is just the marginal distribution from above formula, namely

$$P(X_n = j+1) = P(N_1^{(n-1)} = j) = \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{\substack{n_2, \dots, n_k \in [n-1] \\ j + \sum_{i=2}^{n} n_i = n-1}} \frac{1}{j \cdot n_2 \cdots n_k}. \tag{5}$$

If $j = n-1$, (5) equals $\frac{1}{n-1}$. For $1 \le j < n-1$, we have

$$P(X_n = j+1) = \sum_{k=2}^{n-j} \frac{1}{k! j} \sum_{\substack{n_2, \dots, n_k \in [n-1-j] \\ \sum_{i=2}^{k} n_i = n-1-j}} \frac{1}{n_2 \cdots n_k}$$
$$= \frac{1}{j} \sum_{k=1}^{n-1-j} \frac{1}{(k+1)!} \sum_{\Delta(n-1-j,k)} \frac{1}{n_1 \cdots n_k},$$

where the last equation is due to an index shift. This shows b).
To show c), we compare (5) with $E((1+K_{n-1-j})^{-1})$. First note that for $j = n-1$, we have $K_0 = 0$ and thus

$$\frac{1}{n-1} E\left(\frac{1}{K_0 + 1}\right) = \frac{1}{n-1},$$

which matches the expression in b). Now assume $1 \le j < n-1$. If we look at the table sizes in exchangeable order, we can compute $P(K_{n-1-j} = k)$ by summing up the probabilities of all possible configurations of table sizes of exactly k

occupied tables in a $CRP(n-1-j)$. Using (4), this leads to

$$E\left(\frac{1}{1+K_{n-1-j}}\right) = \sum_{k=1}^{n-1-j} \frac{1}{k+1} \sum_{\Delta(n-1-j,k)} \frac{1}{k! n_1 \cdots n_k}$$
$$= \sum_{k=1}^{n-1-j} \frac{1}{(k+1)!} \sum_{\Delta(n-1-j,k)} \frac{1}{n_1 \cdots n_k}.$$

Comparison with (5) yields

$$P(X_n = j+1) = \frac{1}{j} E\left(\frac{1}{1+K_{n-1-j}}\right).$$

Recall that K_{n-1-j} is distributed as the number of cycles in a uniform permutation of $[n-1-j]$. It is well-known that the number of cycles is distributed as the sum of independent Bernoulli variables $B_1, \ldots, B_{n-1-j}$ with success probability $\frac{1}{i}$ for B_i (e.g., see [1, p.10]). This proves c). □

Remark. Let K_n be the number of occupied tables in a $CRP(n)$. Using $K_n \stackrel{d}{=} \sum_{i\in[n]} B_i$ for independent Bernoulli variables with success probability $\frac{1}{i}$, we deduce the recursion

$$E\left(\frac{1}{m+K_n}\right) = (1-\frac{1}{i})E\left(\frac{1}{m+K_{n-1}}\right) + \frac{1}{i}E\left(\frac{1}{m+1+K_{n-1}}\right)$$

for all $m \in \mathbb{N}_0$. This recursion gives an efficient method to compute the distribution of the minimal clade size X_n by using the representation in Lemma 2.3 c).

Remark. In [16], Goldschmidt and Martin have proven the weak convergence result for M_n for $n \to \infty$ by using the construction of the Bolthausen-Sznitman n-coalescent via cutting a random recursive tree and embedding the random recursive tree in a Yule process. However, as also hinted at by Goldschmidt and Martin (see [16, Cor. 3.3, Remark a)]), the representation of M_n as a uniformly chosen table in a $CRP(n-1)$ allows to use results about uniform random permutations to prove the convergence part of 2.2 without using the Yule process embedding.

Proof. (Alternative proof of 2.2)
First, let's look at the distribution function of $\frac{\log(X_n-1)}{\log(n-1)}$. Let $x \in [0,1]$. Using Lemma 2.3 a), we get

$$P\left(\frac{\log(X_n-1)}{\log(n-1)} \le x\right) = P(X_n - 1 \le (n-1)^x) = \sum_{j=1}^{\lfloor (n-1)^x \rfloor} E\left(\frac{A_j^{(n-1)}}{K_{n-1}}\right)$$
$$= E\left(\frac{\sum_{j=1}^{\lfloor (n-1)^x \rfloor} A_j^{(n-1)}}{\sum_{i=1}^{n-1} A_i^{(n-1)}}\right), \qquad (6)$$

where $A_i^{(n-1)}$ is the number of tables with exactly i customers in a $CRP(n-1)$. The functional central limit theorem of DeLaurentis and Pittel [11] (see also Hansen [20]) states

$$\left(\frac{\sum_{j=1}^{\lfloor n^x \rfloor} A_j^{(n)} - x \log n}{\sqrt{\log n}}\right)_{x\in[0,1]} \xrightarrow{d} (B_x)_{x\in[0,1]},$$

in $D[0,1]$ when $n \to \infty$, where B is a standard Brownian motion. This implies

$$\frac{\sum_{j=1}^{\lfloor n^x \rfloor} A_j^{(n)}}{\log n} \xrightarrow{p} x$$

for $x \in [0,1]$. We apply this result to both the nominator and denominator of the right hand side of (6) (inside of $E(\cdot)$) and get

$$\frac{\sum_{j=1}^{\lfloor (n-1)^x \rfloor} A_j^{(n-1)}}{\sum_{i=1}^{n-1} A_i^{(n-1)}} = \frac{\sum_{j=1}^{\lfloor (n-1)^x \rfloor} A_j^{(n-1)}}{\log(n-1)} \frac{\log(n-1)}{\sum_{i=1}^{n-1} A_i^{(n-1)}} \xrightarrow{p} x$$

for $n \to \infty$. Since $0 \le \frac{\sum_{j=1}^{\lfloor (n-1)^x \rfloor} A_j^{(n-1)}}{\sum_{i=1}^{n-1} A_i^{n-1}} \le 1$ for all x, n, we have uniform integrability and hence

$$E\left(\frac{\sum_{j=1}^{\lfloor (n-1)^x \rfloor} A_j^{(n-1)}}{\sum_{i=1}^{n-1} A_i^{(n-1)}}\right) \to x$$

for $n \to \infty$ which shows

$$\frac{\log(X_n - 1)}{\log(n-1)} \xrightarrow{d} U_{[0,1]},$$

where $U_{[0,1]}$ is the uniform distribution on $[0,1]$. $\frac{\log(X_n)}{\log n}$ behaves in the same way which can be shown with a Slutski argument. □

For the asymptotics of moments of X_n (as well as M_n and RT_n), we use the expression for P_{X_n} from Lemma 2.3 c), namely

$$P(X_n = j+1) = \frac{1}{j} E\left(\frac{1}{1+K_{n-1-j}}\right),$$

where K_n is the number of occupied tables in a $CRP(n)$, we will be ab. Note that K_n also gives the number of cycles in a uniform permutation of $\{1,\dots,n\}$. Thus, the distribution of K_n is given by

$$P(K_n = k) = \frac{S_{n,k}}{n!} \quad \text{for } k \in [n], \tag{7}$$

where $(S_{n,k})_{k\in[n],n\in\mathbb{N}}$ denote the absolute Stirling numbers of the first kind. It is well-known that (see, e.g., [23, Eq. 3.2])

$$\frac{K_n}{\log n} \to 1 \quad \text{almost surely} \tag{8}$$

for $n \to \infty$. Since we want to use Lemma 2.3 c), we're more interested in the behaviour of $E((1+K_n)^{-1})$. From (8), we immediately get

$$\frac{\log n}{1+K_n} \to 1 \quad \text{almost surely} \tag{9}$$

for $n \to \infty$. We will need a L^1-version of (9).

Lemma 2.4.

$$\frac{\log n}{1+K_n} \to 1 \quad \textit{in } L^1 \textit{ for } n \to \infty.$$

Proof. The result follows from (9) and the uniform integrability of $\frac{\log n}{1+K_n}$, which we show now. Note that since $\frac{\log n}{1+K_n} \leq \frac{\log n}{K_n}$ for all $n \in \mathbb{N}$, it suffices to show uniform integrability for $\frac{\log n}{K_n}$. Let $A > 0$ and $H_n \overset{d}{=} Pn(\log n)$ be a Poisson-distributed random variable with parameter $\log n$. Note that $H_n \overset{d}{=} \sum_{i \in [\log n]} H_i^{(1)}$, where $(H_i^{(1)})_{i \in \mathbb{N}}$ are i.i.d. with $H_1^{(1)} \overset{d}{=} Pn(1)$. For $A > 1$, we have

$$\begin{aligned}
\int_{\left\{\frac{\log n}{K_n} \geq A\right\}} \left|\frac{\log n}{K_n}\right| dP &= \log n \sum_{k=1}^{A^{-1}\log n} \frac{1}{k} P(K_n = k) \\
&\overset{(7)}{=} \log n \sum_{k=1}^{A^{-1}\log n} \frac{S_{n,k}}{n!k} \\
&= \sum_{k=1}^{A^{-1}\log n} \frac{(\log n)^k}{k!} e^{-\log n} \left(\frac{1}{\Gamma(1+r)} + O\left(\frac{k}{(\log n)^2}\right)\right) \\
&\leq CP\left(H_n \leq \frac{\log n}{A}\right) \\
&= P\left(\frac{\sum_{i \in \log n} H_i^{(1)}}{\log n} \leq A^{-1}\right) \to 0
\end{aligned}$$

for $n \to \infty$, where $r = (k-1)(\log n)^{-1}$ and C is a suitable constant. Here, we use the uniform asymptotic expansion from Hwang (see Theorem 2 in [18]) for the absolute Stirling numbers $S_{n,k}$ of the first kind for $1 \leq k \leq A^{-1}\log n$ (we actually use the cruder version from [1, Eq. 1.30]). The convergence to 0 follows from the law of large numbers for $(H_i^{(1)})_{i \in \mathbb{N}}$.
This computation shows the uniform integrability of $\frac{\log n}{1+K_n}$ and thus the lemma. □

Theorem 2.5. *For $n \in \{2, 3, ...\}$, let X_n be the minimal clade size in the Bolthausen-Sznitman n-coalescent. For all $k \in \mathbb{N}$, we have*

$$\frac{\log n}{n^k} E(X_n^k) \to \frac{1}{k}$$

for $n \to \infty$.

Again, this theorem is also true for M_n and RT_n instead of X_n.

Proof. Using Lemma 2.3 c), we get

$$\begin{aligned} E((X_n-1)^k) &= \sum_{j=1}^{n-1} j^{k-1} E\left(\frac{1}{1+K_{n-1-j}}\right) \\ &= \sum_{l=0}^{n-2} (n-1-l)^{k-1} E\left(\frac{1}{1+K_l}\right) \\ &= \sum_{i=0}^{k-1} \binom{k-1}{i} (n-1)^{k-1-i} (-1)^i \sum_{l=0}^{n-2} l^i E\left(\frac{1}{1+K_l}\right) \end{aligned}$$

We will now use Karamata's Tauberian theorem for power series (see [3, Corr. 1.7.3]). It states (among other things) that if $a_l \sim \frac{c}{\Gamma(\rho)} l^{\rho-1}\mathcal{L}(l)$ for $n \to \infty$, where $c, \rho > 0$ and $\mathcal{L}$ is a slowly varying function, then $\sum_{k\in[n]} a_k \sim \frac{c}{\Gamma(1+\rho)} n^\rho \mathcal{L}(n)$. We define $a_l := l^i E\left(\frac{1}{1+K_l}\right)$. Note that $a_l \sim \frac{l^i}{\log l}$ for $l \to \infty$ due to Lemma 2.4, which enables us to use the Tauberian theorem for a_l with $c := \Gamma(i+1) = i!$, $\rho = i+1$ and $\mathcal{L}(n) = (\log n)^{-1}$. For each $i \in [k-1]_0$, we thus have

$$\sum_{l=0}^{n-2} l^i E\left(\frac{1}{1+K_l}\right) \sim \frac{1}{i+1}\frac{n^{i+1}}{\log n}$$

for $n \to \infty$. This shows

$$\begin{aligned} \frac{\log n}{n^k} E((X_n-1)^k) &= \sum_{i=0}^{k-1} \binom{k-1}{i} (n-1)^{k-1-i} \frac{\log n}{n^k} (-1)^i \sum_{l=0}^{n-2} l^i E\left(\frac{1}{1+K_l}\right) \\ &\sim \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{\log n}{n^{i+1}} (-1)^i \frac{1}{i+1}\frac{n^{i+1}}{\log n} \\ &= \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{(-1)^i}{i+1} = \frac{1}{k} \end{aligned}$$

for $n \to \infty$, where the last equation follows by elementary calculations. Thus, for each $k \in \mathbb{N}$, we have established

$$\frac{\log n}{n^k} E((X_n-1)^k) \to \frac{1}{k}$$

for $n \to \infty$. The theorem now is proven as

$$\frac{\log n}{n^k} E(X_n^k) = \sum_{i\in[k]_0} \binom{k}{i} \frac{\log n}{n^k} E((X_n-1)^i) \to \frac{1}{k}$$

for $n \to \infty$ □

Remark. This last result fits well with the notion that X_n can heuristically be seen as n^U when n is big (U uniformly distributed on $[0,1]$) following from Theorem 2.2, since the kth moment of n^U is $\frac{n^k}{k \log n}$.
We compare this heuristic to the results for X_n in Kingman's n-coalescent from [6, p.4], where the authors state that X_n, without scaling, converges to a Yule distribution of parameter $\rho = 2$. So in the Bolthausen-Sznitman n-coalescent, the minimal clade size is much bigger than in Kingman's coalescent. This agrees with the more starlike shape of a non-Kingman n-coalescent compared to Kingman's n-coalescent.

Acknowledgement. We thank an anymonous referee to point out the connection of the minimal clade with the non-relatives of 1 in the last coalescence event of the Bolthausen-Sznitman n-coalescent and hence the proof of the convergence part of Theorem 2.2. We thank J.-S. Dhersin, M. Möhle and L. Yuan for a fruitful discussion concerning the alternative proof of Theorem 2.2 and also E. Teufl for advice concerning which journal to submit to.

References

[1] Arratia, R., Barbour, A. D. and Tavaré, S. (2003) *Logarithmic Combinatorial Structures: A Probabilistic Approach* EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich.

[2] Berestycki, J., Berestycki, N. and Schweinsberg, J. (2012) The genealogy of branching Brownian motion with absorption. To appear in *Ann. Probab.*.

[3] Bingham, N. H., Goldie, C.M. and Teugels, J. L. (1987) Regular variation. *Encyclopedia of Mathematics and its Applications* **27**. Cambridge University Press, Cambridge. MR898871.

[4] Bolthausen, E. and Sznitman, A.-S. (1998) On Ruelle's probability cascades and an abstract cavity method. *Commun. Math. Phys.* **197**, 247–276. MR1652734.

[5] Bovier, A. and Kurkova, I. (2007) Much ado about Derrida's GREM. In: *Spin Glasses*, *Lecture Notes in Math.*, **1900**, Springer, Berlin, 81–115. MR2309599.

[6] Blum, M.G.B. and François, O. (2005) Minimal clade size and external branch length under the neutral coalescent. *Adv. Appl. Probab.* **37**, 647–662. MR2156553.

[7] Brunet, É and Derrida, B. (2012) Genealogies in simple models of evolution. *Arxiv preprint* arXiv:1202.5997.

[8] Brunet, É, Derrida, B., Mueller, A.H. and Munier, S. (2006) Noisy traveling waves: effect of selection on genealogies.. *Europhys. Lett.* **76**, 1–7.

[9] Brunet, É, Derrida, B., Mueller, A.H. and Munier, S. (2007) Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Phys. Rev. E* **76**, 041104.

[10] Caliebe, A., Neininger, R., Krawczak, M. und Rösler, U. (2007) On the length distribution of external branches in coalescence trees: genetic diversity within species. *Theor. Popul. Biol.* **72**, 245–252.

[11] DeLaurentis, J.M. and Pittel, B. (1985) Random permutations and Brownian motion. *Pac. J. Math.* **119**, 287–301.

[12] Desai, M.M., Walczak, A.M. and Fisher, D.S. (2012) Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, published online before print.

[13] Dhersin, J.S., Freund, F., Siri-Jégousse, A. and Yuan, L. (2012) On the length of an external branch in the Beta-coalescent. To appear in *Stoch. Proc. Appl.* Preprint available via arXiv:1201.3983

[14] Freund, F. and Möhle, M. (2009) On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent. *Markov Process. Related Fields* **15**, 387–416. MR2554368.

[15] Y.-X. Fu and W.-H. Li. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709, 1993.

[16] Goldschmidt, C. and Martin, J.B. (2005) Random recursive trees and the Bolthausen-Sznitman coalescent. *Electron. J. Probab.* **10**, 718–745. MR2164028

[17] Gnedin, A., Iksanov, A. und Möhle, M. (2008) On asymptotics of exchangeable coalescents with multiple collisions. *J. Appl. Probab.* **45**, 1186–1195. MR2484170.

[18] Hwang, H.-K. (1995). Asymptotic expansions for the Stirling numbers of the first kind. *J. Combin Theory Ser. A* **71(2)**, 343–351. MR1342456.

[19] Hallatschek, O. and Neher, R.-H. (2012) Genealogies of rapidly adapting populations. Will be published in *PNAS*. Available also as *Arxiv preprint* arXiv:1208.3185.

[20] Hansen, J.C. (1990) A functional central limit theorem for the Ewens sampling formula. *J. Appl. Probab.* **27**, 28–43.

[21] J.F.C. Kingman. The coalescent. *Stoch. Proc. Appl.* 13: 235–248, 1982.

[22] Pitman, J. (1999) Coalescents with multiple collisions. *Ann. Probab.* **27**, 1870–1902. MR1742892.

[23] Pitman, J. (2005) Combinatorial Stochastic Processes. In: *Ecole d'Et der Probabilités de Saint-Flour XXXII–2002*, Editor: Picard, J., *Lecture Notes in Mathematics*, **1875**, Springer. MR2245368.

[24] Sagitov, S. (1999) The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36**, 1116–1125. MR1742154.

Fabian Freund
Crop Plant Biodiversity and Breeding Informatics Group (350b), Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Fruwirthstrasse 21, 70599 Stuttgart, Germany
Email address: fabian.freund@uni-hohenheim.de

Arno Siri-Jégousse
CIMAT, A.C., Calle Jalisco s/n, Col. Mineral de Valenciana, 36240 Guanajuato, Guanajuato, Mexico
Email address: arno@cimat.mx